Paper prepared for the 23rd Conference of Australian Institutes of Transportation Research, Monash University, December 10 –12, 2001

LOCAL SAMPLE UPDATES FOR SYNTHETIC HOUSEHOLD TRAVEL SURVEY DATA

Stephen P. Greaves,

Institute of Transport Studies, Department of Civil Engineering,
Monash University, VIC 3800, Australia.
Stephen.Greaves@eng.monash.edu.au

ABSTRACT

This paper reports on a continuing research effort to refine and improve the simulation of "synthetic" household travel survey data for regions that do not have the resources to collect such data. Original testing of the approach suggested that while the approach was able to reproduce trip rates that were generally comparable at an aggregate and disaggregate level, other salient trip characteristics (particularly mode and trip length) were less effectively replicated. Presented here is a refinement to the process in which data from a small local sample are used to update these synthetic data. Initial results suggest this procedure is capable of generating a travel survey data set that is more reflective of local conditions.

Keywords: simulation, synthetic data, data updating.

1. INTRODUCTION

This paper reports on a continuing research effort to refine and improve the simulation of "synthetic" household travel survey data for regions that do not have the resources to collect such data. The rationale of the approach was first described in *[1]* but for the benefit of the reader is summarised here. Households in a travel survey database are classified into relatively homogeneous sociodemographic groups for each travel attribute of interest[1]. Within each group, this attribute will vary. This variation is captured in an empirical distribution that (effectively) forms the basis for simulating that attribute. Households in a region (with no travel data) are then classified into the same categories as those used in the original segmentation. A simulation procedure is then used to sample from the relevant distributions and the value is assigned to each household. This process is repeated for each attribute of interest producing a full synthetic travel record for each household.

Initial results and analysis proved this was a concept worth pursuing further *[2]*. The next step was to test the wider applicability of the approach using urban areas of differing characteristics that had conducted recent activity/travel surveys. The results of this comparison are reported on in these conference proceedings *[3]*. While the results were generally encouraging it was clear that differences in some of the travel characteristics (particularly mode and trip length) were only partially captured through the demographic segmentation of the population. This was not an unexpected finding because logically these attributes are related to characteristics of the region (e.g., area, public transportation coverage) in addition to characteristics of the individual and their household.

[1] The data simulated are trip rates, then for each trip the mode, departure time and trip length (minutes).

With this in mind, the current challenge is to "localise" the synthetic travel data to account more effectively for regional differences. This paper suggests one potential method in which the synthetic data set is updated using a small sample of local data collected from households in the application region. Following an explanation of the rationale for such an approach the paper reports on some preliminary tests of its effectiveness that show the approach has merit. Finally, some critical issues and possible future research directions are outlined.

2. BACKGROUND/RATIONALE

The most logical approach to localise the synthetic data is to generate the simulated travel data from regions that are similar to the application region. This similarity could be measured in terms of classifiers such as population size, density, public transportation coverage etc. This is the typical approach taken in selecting regions for borrowing travel-demand models, a practice that has met with mixed results *[4]*. This appears to be due to various factors including unknown differences between model parameters in the estimation and application contexts (the transfer bias), specification errors in the original model and a lack of consensus over how to measure and evaluate the transfer "success."

The use of data from regions of similar characteristics produced marginal improvements during the original development of the synthetic data method *[2]*. The next stage was to examine the potential for improving the process by updating the synthetic data itself with data collected in the application region. The rationale for this idea came from the consistent improvements reported in the literature with respect to the use of local data to update borrowed model parameters *[5]*. Such data might include aggregate transportation measures (e.g., mode shares) that are used to update model constants or more preferably a small disaggregate sample of household travel survey data. In this case, these data can be used to update model parameters using various methods whose effectiveness is seemingly dependent on the size of the update sample, the specification error in the original model and the transfer bias *[5, 6]*.

While this evidence is presented in the context of improving model transfer, no conceptual reason is apparent why this idea could not be applied to *update travel data.* The idea of updating travel data has received little attention to date. In the only study of this topic of which the author is aware, a small sample of Baton Rouge households (108 observations) are used to update aggregate trip rates, mode shares, and trip lengths from the 1995 Nationwide Personal Transportation Survey (NPTS) for the Baton Rouge metropolitan area using a Bayesian updating procedure *[7]*. While the sample was too small to provide stable results, the use of a simulated sample of 450 households (developed from urban sections of regions of similar population sizes) provided aggregate measures of transferred data comparable to locally-collected data.

This limited evidence suggests that data from a small sample could be used to update the synthetic data to be more reflective of local conditions. While the simplest approach would be to update trip rates, mode shares, departure times and trip lengths after the fact this would lose the conditionality that is a feature of the simulation process. Each step builds on the previous step with the mode dependent on the simulated trip purpose, the departure time dependent on the simulated mode (and consequently the purpose), and the trip length dependent on the simulated departure time (and consequently the mode and purpose). The implication is that the update process must be incorporated into the simulation process itself.

The approach taken here is to focus on the driving force behind the simulation, namely the *empirical distributions* of the trip attributes. If the local sample was segmented into the same categories as those used in the original procedure then distributions of attributes could be developed as before. These local distributions could then be combined with the original distributions to develop a set of updated distributions that could form the basis for sampling in the simulation process.

3. SYNTHETIC DATA UPDATING APPROACH

With this in mind, an approach is conceptualised and tested here to update the synthetic data with local data from a small disaggregate sample of households. In the original work *[2]*, travel data were simulated for households from the 1997 Baton Rouge Personal Transportation Survey (BRPTS) using travel data from the 1995 NPTS[2]. Comparisons were then run against the actual travel data reported in the BRPTS. For the purposes of the current test, a sub-sample of 200 households from the BRPTS was taken to form the update sample. Comparisons were made against the remaining 784 households to avoid the bias caused by having the comparison sample partially composed of the update sample.

Households in the update sample were classified into the same categories as those used in the original procedure. Distributions of each attribute were developed as before. These distributions were then used to update the original distributions using a Bayesian updating procedure – similar procedures have been used in model updating *[5]* and to update aggregate data values in the work reported in *[7]*.

Under this procedure, an unknown parameter, θ is related to its prior distribution and the likelihood function of the local data by the probability expression:

$$\begin{pmatrix}\text{Posterior probability}\\ \text{of } \theta \text{ given the local data}\end{pmatrix} \propto \begin{pmatrix}\text{prior probability}\\ \text{of } \theta\end{pmatrix} * \begin{pmatrix}\text{likelihood function of}\\ \text{the local data given } \theta\end{pmatrix} \qquad (1)$$

The critical issue with using Bayesian updating is to define the prior distribution of θ. The most widely used approach is to assume θ is normally distributed with mean θ_t and variance, σ_t. Similarly the sampling distribution of the local data is assumed to be normally distributed with mean θ_s and variance, σ_s. This assumption (conjugate prior) enables data from the two sources to be combined to produce a posterior distribution that is also normally distributed with parameters θ_p and variance, σ_p that are calculated as follows:

$$\theta_p = \left[\theta_t / \sigma_t^2 + \theta_s / \sigma_s^2\right] / \left[1/\sigma_t^2 + 1/\sigma_s^2\right] \qquad (2)$$

$$\sigma_p^2 = \left[1/\sigma_t^2 + 1/\sigma_s^2\right] \qquad (3)$$

Equation 2 shows that θ_p is derived from the prior and local sample, which have effectively been weighted by the inverse of their respective variances. These weights can be manually altered if it is deemed that the update sample should be given more impact in the updating procedure. As a

[2] The NPTS is a national survey of 42,033 households. The sample is stratified across each day of the year and various spatial criteria. The BRPTS was conducted using the same methods as the NPTS and comprised a sample of 1,395 households drawn from the Baton Rouge metropolitan region.

practical matter, the discrepancy in sample size between the update sample and the NPTS sample meant that without some manual adjustment of weights, minimal effects were observed. This is clearly a matter for contention and a review of the literature suggests this comes down to sound reasoning rather than clearly defined rules. In this case, the variance was weighted by the proportion of trips in the BRPTS to the NPTS for each category. Where less than five trips were observed in the BRPTS, the original distribution from the NPTS was taken.

Given that each interval was treated as a proportion, an estimate is needed for the standard error of the share – this is analogous to the standard deviation of the sampling distribution of a sample proportion. This can be derived from the following expression although it must be noted this requires five or more estimates for the assumption of normality to hold. This was problematic given the size of the update sample and the level of disaggregation used.

$$std.error = \frac{\sqrt{p(1-p)}}{n} \text{ where the sample proportion } \frac{x}{n} \text{ is substituted for } p.$$

$x = \text{share}, n = \text{sample size}$

3.1 An Example

To illustrate how this procedure works, consider the case of simulating travel mode for a target region. The objective is to predict the mode of travel (Privately Occupied Vehicle-driver, Privately Occupied Vehicle -passenger, public transport, bike/walk) for each simulated trip. Households (and their trips) in the NPTS are classified into the 39 categories shown in Table 1. The next stage is to develop the sampling distributions for mode for each category based on their frequency of occurrence in the NPTS database. For instance, for households falling in the "Home-Work, 0 Vehicles" category, 12% of trips are by POV-driver, 34% by POV-passenger, 37% by public transport and 17% by Bike/Walk.

Table 1
Categorization Scheme for Mode Simulation

Trip Purpose	**Mode Categories**
Home-Work	1 = 0 Vehicles, 2 = 1 Vehicle, 1 Worker, 3 = 1 Vehicle, 2+ Workers, 4 = 2+ Vehicles, 1-2 Workers, 5 = 2+ Vehicles, 3+ Workers
Home-School	6 = 0 Vehicles, 7 = 1 Vehicle, 1-3 Persons, 8 = 1 Vehicle, 4+ Persons, 9 = 2 Vehicles, 1-2 Persons 10 = 2 Vehicles, 3 Persons, 11 = 2 Vehicles, 4 Persons, 12 = 2 Vehicles, 5+ Persons, 13 = 3+ Vehicles, 1-3 Persons, 14 = 3+ Vehicles, 4+ Persons
Home-College	15 = 0 Vehicles, 16 = 1 Vehicle, 0-1 Persons Aged 18-24, 17 = 1 Vehicle, 2+ Persons, 18 = 2+ Vehicles, 2+ Persons Aged 18-24, 19 = 2 Vehicles, 1-2 Persons Aged 18-24, 20 = 2+ Vehicles, 3+ Persons Aged 18-24, 21 = 3+ Vehicles
Home-Shop	22 = 0 Vehicles, 23 = 1+ Vehicle, 1 Person, 24 = 1 Vehicle, 2+ Persons, 0 Children (5-17) 25 = 2+ Vehicles, 2+ Persons, 0 Children (5-17), 26 = 1+ Vehicle, 1+ Children (5-17) 27 = 2+ Vehicles, 1 Child (5-17), 28 = 1+ Vehicle, 2+ Children (5-17)
Home-Other & Non-Home-Other	29 = 0 Vehicles 1+ Vehicle, 30 = 1 Person (18+), 0 Children (5-17), 31 = 1+ Vehicle, 2+ Persons (18+), 0 Children (5-17), 32 = 1 Vehicle, 1 Child (5-17), 33 = 2+ Vehicles, 1 Child (5-17), 34 = 1 Vehicle, 2+ Children (5-17), 35 = 2+ Vehicles, 2+ Children (5-17)
Work-Other	36 = 0 Vehicles, 37 = 1 Vehicle, 1 Worker, 38 = 1 Vehicle, 2+ Workers, 39 =2+ Vehicles

For the update sample of 200 households, the distributions within this same category are 37% of trips by POV-driver, 63% by POV-passenger, 0% by public transport and 0% by bike/walk. This distribution is then used to update the original distribution to give the updated values of 22% of trips by POV-driver, 60% by POV-passenger, 7% by public transport and 11% by bike/walk – this is the new distribution that will be used in the simulation procedure.

Table 2 shows the impacts of updating on the distributions for the simulation of mode for the five home-work categories. This illustrates a potential problem of working at a disaggregate level with small samples. For instance, no public transport trips were captured in the update sample and two of the categories captured less than 10 trips. This raises the issue of whether the update sample should be drawn by pre-defined stratum to ensure adequate representation of households in these categories. This issue is discussed in a later section and is currently being investigated.

Table 2
Mode Share Distributions for the Home-Work Categories

Category	**Data Source**	**No. of Trips**	**Driver**	**Passenger**	**Transit**	**Bike/Walk**
0 Vehicles	Update Sample	8	37%	63%	0	0
	NPTS**	91	12%	34%	37%	17%
	Updated Distribution	91	22%	60%	7%	11%
1 Worker, 1 Vehicle	Update Sample	51	100%	0	0	0
	NPTS	623	91%	7%	1%	2%
	Updated Distribution	623	97%	1%	1%	2%
2+ Workers, 1 Vehicle	Update Sample	9	78%	22%	0	0
	NPTS	431	65%	25%	6%	4%
	Updated Distribution	431	67%	23%	5%	5%
1-2 Workers, 2+ Vehicles	Update Sample	212	96%	3%	0	1%
	NPTS	3664	95%	4%	0	1%

	Updated Distribution	3664	95%	4%	0%	1%
3+ Workers, 2+ Vehicles	Update Sample	70	87%	13%	0	0%
	NPTS	1048	87%	9%	1%	2%
	Updated Distribution	1048	87%	12%	0%	1%

**Remaining 784 BRPTS households.*
***Drawn from Regions of 500,000 – 1,000,000 Population*

4. RESULTS

Table 3 shows the mode share comparisons by selected trip purposes using the original distributions and the updated distributions. The overall impression is that the use of the updated distributions has significantly improved the prediction of mode choice for the BRPTS sample across all purpose/mode combinations. While statistically significant differences remain between the predicted and actual shares, this proves that the method has *potential* for gains in the quality of data generated by the simulation process.

Table 3
Mode Share Comparisons

Trip Purpose	Mode	BRPTS (784 households)	Simulated Data (Original Distributions)	Simulated Data (Updated Distributions)
Home-Work	Auto Driver	92.1%	88.6%**	90.8%
	Auto Pass.	6.4%	7.6%	7.2%
	Transit	0.6%	2.1%**	0.6%
	Bike/Walk	0.9%	1.7%	1.4%*
Home-Shop	Auto Driver	77.3%	69.5%**	75.1%
	Auto Pass.	19.6%	23.7%*	21.0%
	Transit	0.1%	1.8%**	0.6%
	Bike/Walk	3.0%	5.1%*	3.3%
Home-Other	Auto Driver	65.2	60.9%**	62.9%
	Auto Pass.	28.6%	31.9%**	29.9%
	Transit	1.0%	2.3%**	1.9%*
	Bike/Walk	5.2%	4.9%	5.4%
Other-Other	Auto Driver	65.8%	60.1%**	63.7%
	Auto Pass.	29.3%	31.9%	29.7%
	Transit	1.8%	1.7%	1.8%
	Bike/Walk	3.1%	6.3%**	4.8%*
All Purposes	Auto Driver	70.9%	66.0%**	68.0%**
	Auto Pass.	22.1%	24.2%**	23.4%*
	Transit	3.3%	4.9%**	4.4%**
	Bike/Walk	3.7%	5.0%**	4.2%

* *Statistically significant at the 95th Percentile Confidence Limit*
** *Statistically significant at the 99th Percentile Confidence Limit*

While it is important to show the procedure produces aggregate mode shares, the updating procedure must also be validated at a disaggregate level for the following reasons. First, one must be wary of potential aggregation bias and how this can create misleading conclusions about whether the procedure is working correctly. Second, problems in this step of the simulation will be propagated through the remaining steps making it imperative they are detected early. Third, one can identify

"problem" segments where the simulation still appears to perform poorly.

Table 4 provides a comparison by number of household vehicles. The results show that the improvements with the updated data while still apparent are less marked than in the previous comparison. In the case of zero vehicle households the updated distributions actually produced poorer results than the original distributions. However, closer inspection and repeated runs of the simulation showed a great fluctuation in results for this comparison because of the small representation of zero-vehicle households in the update sample. This was true for other disaggregate comparisons (not reported here) of certain segments and is discussed further in the next section.

Table 4
Mode Share Comparison by Number of Household Vehicles

Trip Purpose	Vehicles	Data Source	Driver	Passenger	Transit	Bike/Walk
Home-Work	0	BRPTS	41.9%	35.5%	12.9%	9.7%
		Original Distributions	5.4%**	16.2%	51.4%**	27.0%
		Updated Distributions	21.6%	67.6%**	2.7%^{IR}	8.1%^{IR}
	1	BRPTS	90.8%	6.3%	1.1%	1.7%
		Original Distributions	79.9%**	15.3%**	2.0%^{IR}	2.8%^{IR}
		Updated Distributions	87.6%	8.8%	1.6%^{IR}	2.0%^{IR}
	2+	BRPTS	93.9%	5.5%	0.1%	0.5%
		Original Distributions	93.2%	5.6%	0.5%^{IR}	0.7%
		Updated Distributions	91.9%	6.2%	0.3%^{IR}	1.6%**
Home-Other	0	BRPTS	10.8%	44.1%	15.3%	29.7%
		Original Distributions	11.2%	51.2%	19.5%	18.0%*
		Updated Distributions	30.2%**	39.0%	15.1%	15.6%**
	1	BRPTS	68.7%	26.7%	0.4%	4.1%
		Original Distributions	68.8%	25.5%	1.1%^{IR}	4.6%
		Updated Distributions	69.2%	25.5%	0.4%^{IR}	4.9%
	2+	BRPTS	67.5%	28.1%	0.4%	3.9%
		Original Distributions	63.7%**	31.7%**	0.9%*	3.7%
		Updated Distributions	64.3%*	31.9%**	0.6%	3.3%
All Purposes	0	BRPTS	15.6%	38.3%	19.3%	26.8%
		Original Distributions	12.3%	42.5%	23.6%	21.7%
		Updated Distributions	26.5%**	38.5%	17.7%	17.3%**
	1	BRPTS	71.6%	22.0%	1.7%	4.7%
		Original Distributions	68.6%	21.9%	4.1%**	5.4%
		Updated Distributions	72.3%	20.4%	2.9%*	4.3%
	2+	BRPTS	73.3%	21.3%	3.0%	2.4%
		Original Distributions	68.9%**	23.5%**	3.7%**	3.8%**
		Updated Distributions	70.0%**	23.5%**	3.6%	2.9%

**Statistically significant difference in trip rates at the 95th percentile confidence level*
***Statistically significant difference in trip rates at the 99th percentile confidence level*
IR – Insufficient records for the z-test of proportions.

5. DISCUSSION

Local data updating clearly has the potential to improve the quality of borrowed and synthetic data. However, a number of issues were identified during this preliminary work that must be addressed in future applications.

1. The size of the update sample is clearly critical. In the model updating literature, update samples ranging from 200 to 500 households are cited. In dollar amounts (assuming a per household cost of AU$200) this translates to AU$40,000 to AU$100,000. In practice the size

of the sample is likely to come down to a trade-off between available funds and perceived benefits in terms of accuracy. This benefit/cost must be quantified through further investigation.

2. Of even more significance (arguably) than the sample size is the way the sample is drawn. In this application, the sample was drawn randomly. However, other testing in which the sample was drawn using pre-defined stratum showed this had a significant impact on results. The reason for this relates to how the makeup of the update sample transfers through to the categories used in the development of the distributions. For instance the low representation of zero-vehicle households in the definition of the distributions was a major reason why the prediction of mode for these households was poor.

3. The Bayesian updating procedure allows the user to employ subjective judgment on the impact of the update sample through the adjustment of weights. This was necessary for the updating to have any impact but must clearly be based on sound reasoning.

4. The approach should be extended to other trip attributes and other regions. This is the subject of concurrent work.

5. One possible approach to data updating is presented here. It must be established if this is the preferred approach.

6. CONCLUSIONS

Local data updating offers an added dimension to the borrowing and synthetic generation of travel data. In addition to improving the quality of synthetic data, it offers (potentially) a means for regions to keep their travel inventories current at a fraction of the cost of conducting large, expensive surveys. This could also have potential benefits for small-scale planning studies that require an inexpensive means to develop a travel database. While further testing and refinement is needed and the approach must be extended to the simulation of other travel attributes, the method presented here and the preliminary results suggest the approach has substantial merit. In particular, issues relating to the size, structure and weight given to the update sample must be investigated further.

7. REFERENCES

1. Greaves, S.P and Stopher, P.R. (2000) Creating a Synthetic Household Travel/Activity Survey – Rationale and Feasibility Analysis. *Transportation Research Record No. 1706, pp. 82-91.*

2. Greaves, S.P. (2000) Simulating Household Travel Survey Data for Metropolitan Areas. Unpublished Ph.D. Dissertation, Department of Civil and Environmental Engineering, Louisiana State University, Baton Rouge, Louisiana.

3. Stopher, P.R., Greaves, S.P., Kothuri, S., and Bullock, P. (2001) Synthesizing Household Travel Survey Data: Application to Two Urban Areas. *Proceedings of the 23rd Conference of Australian Institutes of Transportation Research, Monash University, December, 2001.*

4. Wilmot, C.G. (1995) Evidence on Transferability of Trip-Generation Models. *Journal of Transportation Engineering, Vo. 121, No. 5, pp. 405-410.*

5. Koppelman, F.S., Kuah, G-K. and Wilmot, C.G. (1985) Transfer Model Updating with Disaggregate Data. *Transportation Research Record No. 1037, pp. 102-107.*

6. Karasmaa, N. and Pursula, M. (1997) Empirical Studies on the Transferability of the Helsinki Metropolitan Area Travel Forecasting Models. *Paper presented at the 76th Annual Meeting of the Transportation Research Board*, Washington DC.

7. Wilmot, C.G. and Stopher, P.R. (2000) Cost-Effective Data Collection in Louisiana. *Report No. 337 prepared for the Louisiana Transportation Research Center.*